Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150007-5

Machine Techniques for Information Selection

Symposium Sponsored by Industrial Liaison Office, Massachusetts Institute of Technology

June 10 and 11, 1952, Cambridge, Mass.

During the past few years, MIT has been investigating the application of machines to the searching and correlating of information. The work has been under the leadership of James W. Perry and has been concerned principally with the problem of analyzing information and designing indexes that can be searched by machine. At the same time, International Business Machines has been developing new machines which will have appropriate operating characteristics.

REPRINTED FROM CHEMICAL AND ENGINEERING NEWS *the newsmagazine of the chemical world*

Vol. 30, pp. 2781, 2806–2810, July 7, 1952

Mechanized System Launches New Era for Literature Searching

New IBM system will require language engineering to exploit its potentialities

Ultra-high speed electronic scanners predicted

Lift for Literature

The bugbear of an exponentially expanding literature seems about to be beaten. Now, for the first time, a completely mechanized system of searching the literature appears to be within the realm of possibility. The new IBM information searching system solves the first problem—identifying articles or abstracts that contain desired information. Other machines are available or can be developed that will take it from there. The Bush Rapid Selector, for instance, in appropriately modified form could be used to make photographic copies (on microfilm) of desired abstracts. It all adds up to rapid and complete searches in hours rather than days or months. But don't order a system yet—a lot of work must be done on coding and putting the literature in a form machines can use. Best guess: five years.

CAMBRIDGE, MASS. — The tedious, "horse" work that is now involved in searching the literature may be made a thing of the past by a new system of handling information that has been unveiled by International Business Machines. The system, which involves several principles entirely new to IBM methods of operation, promises to put literature searches on an almost completely mechanical basis and to permit large segments of the literature to be scanned in incredibly short periods of time.

The system was described at a symposium on machine techniques for information selection at Massachusetts Institute of Technology here recently. Attending were nearly 100 representatives of government agencies and of companies participating in the institute's industrial liaison program. Also predicted at the symposium was the possibility of designing an electronic scanning machine that would search the literature at the rate of 5 million documents per hour.

While these machines will bring new speed to literature searches, what is even more significant is that they will permit one to adopt the policy of total searching, said Thyllis Williams of the MIT Center for International Studies. A machine doesn't get tired or decide that it has gone far enough when it has reviewed "substantially" all of the material available.

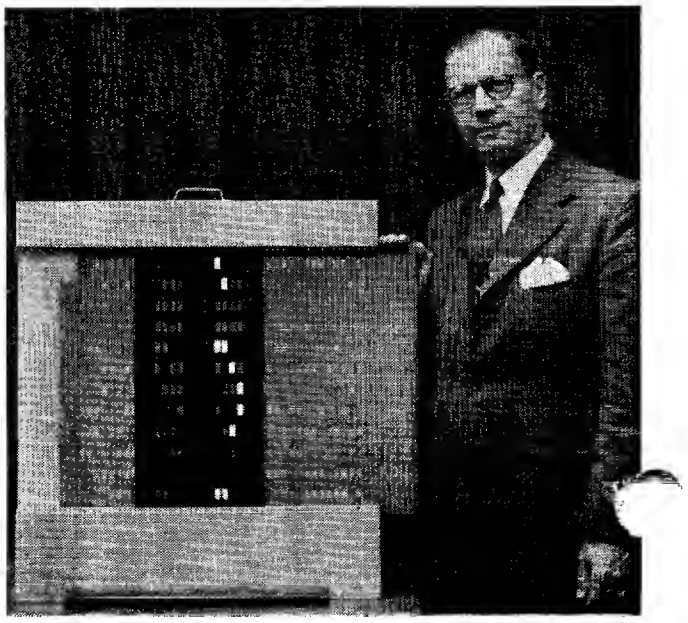
Before machine searching is available for general use, however, a "machine language" will have to be developed. This is over and above the code that is used to express letters, numbers, and symbols in terms of holes in a card; it refers to the way in which the linguistic and numeric elements will be put together so that the machine will understand them. The machine indexing project under James W. Perry at the Center for International Studies is working on this problem.

The IBM Search System. The new system is based on the standard IBM card, but, contrary to customary practice where electrical contact is made by brushes through holes in the card, the new system depends on photoelectric scanning. Further, where position on the card determines the significance of the holes in the usual method of operation, the new system employs patterns of holes that are independent of position.

These facts mean that instead of reserving a portion of the card for certain types of information whether it is present or not, the new system avoids waste of space and permits information to be presented compactly. Coded index information, which can include keys to everything of importance contained in a given document, is punched on the card without reference to order. Thus the information can be said to "float" on the card. And if the amount of information exceeds the capacity of one card, a trailer card or cards can be used. By means of a hold-over command punched at the end of each card for which another follows, the machine will treat the several cards as a continuous record.

A pattern consists of five holes punched out of a column of 12 possibilities. The scanning operation, in which cards containing desired information are selected, is conducted by a matching process. A question card is punched with a pattern that complements that of the answer being sought—seven holes per column punched in the spaces not punched in the answer.

A matching condition—when the answer is superimposed on the question—is one



H. P. Luhn demonstrates IBM information searching system. File card, represented by light panel, carries index information. Question card (dark strip) carries two four-letter words. Word to left is in matching position. Light behind panel shows through holes in both cards for word on right that does not match

in which no light passes. It is a momentary blackout. This can be detected very rapidly by the photocell so the file cards can be passed over the question card in rapid succession—up to 1000 per minute. When a match is found, that card is kicked out.

Each photocell, scanning one to four columns, may be wired to act independently of any other, or several may be wired together to yield such answers as A + B + C, A or B or C, A + B but not C, or any pair among A, B, and C.

In addition to the scanner the system requires three other machines—a card punch, a sorter, and a transcriber. In the card punch, information is punched onto the card in coded form through a typewriter keyboard. By throwing a switch the machine will punch complementary patterns for making question cards. Duplicate cards may be made in standard card reproducing machinery.

In the sorter the cards that have been selected by the scanner are arranged in any desired order. In the transcriber information on the selected cards is automatically typed out, ready for decoding. The transcriber may take all of the information off each card, or it may be rigged to select only certain information, such as serial numbers of literature references in the master file. Alternatively, the actual reference information might be included on the card so that the transcriber yields a list of references directly.

Only one experimental model of each of the new machines has been built to date, and these are now being used for exploratory work on coding systems. The machines are said to be of about the same order of complexity as standard IBM machines, and it is anticipated that when ready they will be available on similar terms. Production prototypes, however, have not yet been made.

The background for this development was laid in the recognition among scientists who handle literature that man's capacity for handling the rapidly increasing volume of technical literature will soon be reached. Two such scientists, Dr. Perry and G. Malcolm Dyson, approached Thomas J. Watson, president of IBM, about four years ago and convinced him that machines could be developed that would solve this problem. Further, they showed Mr. Watson that in the Dyson code they already had a system of notation for organic compounds that is adaptable to machine operations.

Mr. Watson accepted the challenge and assigned the project to H. P. Luhn of the company's engineering laboratory in Poughkeepsie, N. Y. Mr. Luhn started with the standard IBM card, which has 80 vertical columns of 12 spaces each. Adopting the principles mentioned above (pattern matching and photoelectric scanning), Mr. Luhn decided that the card would have to be scanned in the longitudinal direction rather than laterally, as is the customary IBM practice. Thus, the cards would pass through the scanner endwise rather than sidewise so that the pattern would be a vertical one based on 12 positions.

In considering the number of holes to be punched in each column, Mr. Luhn recognized that while six holes would give the largest number of combinations (924), five holes would give groups of patterns that are better suited to the decimal system of numbers and to the English alphabet. And the 792 combinations that are possible by punching five holes out of 12 positions were considered adequate.

The top two rows of holes in the IBM card are usually handled separately as key or index holes (labeled X and Y), leaving 10 in the column below. These 10, numbered 0 to 9, yield combinations totaling 792 in three series, depending on whether X and Y are not punched, one punched, or both punched. In each case, only five holes are punched per column, including X and Y.

Mr. Luhn has devised a scanning code in several series which provide for two full alphabets (each with capitals and small letters), several sets of two-digit numbers, subscripts, and superscripts as well as special characters and typewriter operations. Even with all of these he has not exhausted the possibilities, nor does it seem practical to do so.

To illustrate, in the series where neither X nor Y is punched, the letter a is recorded by punching holes 98765; the letter B is recorded by punching 53210; and the number 14 is represented by 86531. (In this series, tens are coded by punching three holes from 5 through 9, units by two holes from 0 through 4. Thus only one column is required to code numbers 00 through 99.) Therefore only three columns would be used to code the symbols aB14. This points up the compactness by which numerical information can be expressed by this code.

Similar compactness can also be obtained for an alphabetical code by expressing binary letter combinations in terms of numbers. Thus by making two-letter combinations consisting of a consonant (except q) followed by a vowel, 100 combinations result. Each combination can then be expressed in a single column on the IBM card. This system, which is referred to as Luko, is particularly adapted to the building of codes in terms of which index information can be transferred to punched cards.

Machine Language. The machine system will operate on open language, such as English, but in so doing it would operate only inefficiently, according to Thyllis Williams. The difficulty lies in the inability of a machine to recognize relationships between natural words. In systematic "machine language," however, interrelationships would be "built in." The use of open language would require the individual who coded the original document to anticipate all possible future demands upon that document. But the researcher, said Mr. Luhn, is looking for novel relationships between a number of subjects, which could not have been anticipated by the indexer. Clearly, a new effort is called for—an effort termed by Miss Williams as language engineering.

Exploratory work is being conducted at MIT on the development of a machine language on the basis of "abstraction ladders," according to Allen Kent of the Center for International Studies. In abstraction ladders, terms are arranged in accordance to their relation to each other, and symbols such as Luko are assigned in accordance with those relationships. Thus, in a typical abstraction ladder, "animal" would be represented by the Luko symbol, NA, "dog" by NADO.

To punch the code for "dog" on a card, two columns would be required—NA in the first, DO in the second. NA in Luko is represented by the number 50; DO by 13. These are coded by 7810 and 7530, respectively, with either X or Y punched. (X is punched for word start, Y for word continuation.) To code the word for dog, therefore, would require that X7810 be punched in the first column, Y7530 in the second. Then to search for "dog," the complements of these combinations would be punched in the question card, and both columns would be controlled by a single photocell.

If a document spoke of the incidence of rabies in dogs, the indexer might feel that listing under "rabies" and under "dogs" would be sufficient. But if an investigator wanted to use open language to conduct a machine search for the incidence of rabies in animals, he would have to encode the question card with names of all animals. With machine language such as the above, however, a search for NA alone would produce documents pertaining to dogs as well as to all other animals. The abstraction ladder would be especially important in searching patents, where the breadth of claims is often a problem. This has been demonstrated by experimental work under the direction of Benjamin E. Lanham at the U. S. Patent Office.

In the MIT machine coding project scientific and technical terms and their definitions are being gathered on Keysort cards for processing. This processing consists first in assigning a term to one of the five categories: (a) processes, (b) machines, (c) materials, (d) attributes, and (e) abstract concepts. The terms are also categorized according to the general field to which they pertain: (a) chemical, (b) physical, (c) mechanical, (d) biological, (e) general. The card is then punched appropriately for each of these categories.

The subject classes are then broken down still further to facilitate studying relationships between terms. Lists of these terms are being given to specialists in various fields who criticize the groupings and expand the files. Final steps will be

to code the terms and to prepare a code dictionary.

Dictionaries, textbooks, indexes, and catalogs are being used as sources for terms. To date, said Mr. Kent, more than 15,000 terms have been gathered and are in various stages of processing. A few hundred of them are even down to the coding level. An objective in this program, he pointed out, is to make the machine language simple enough so that highly skilled personnel will not be required.

The AMERICAN CHEMICAL SOCIETY has been cooperating in this project through its Committee on Scientific Aids to Literature Searching. Additional volunteers to participate in this program were invited by Dr. Perry, chairman of the committee.

The Literature Problem. The need for machine searching was pinpointed by Dr. Perry, who opened the symposium, by the statement that all information processes are growing at an exponential rate. This means not just technical literature but communication, electronics, transportation, and others. Average period of doubling is about 20 years, he pointed out.

As the literature becomes more voluminous, said Eugene Scott of the Interdepartmental Committee on Scientific Research and Development, U. S. Government, the expense of handling information rises. It will soon come to the point, said Dr. Scott, where one will spend so much time in searching that he won't have time to do any new work. It has been reported that in Sweden one third of the cost of research is for literature searching.

The volume of material to be searched by lawyers is expanding to the point where increasing difficulty is encountered in maintaining the traditional practice of basing decisions on precedents recorded in the legal literature, according to John M. Maguire of the Harvard Law School.

Speaking of other techniques that have been developed to speed communication, Mortimer Taube of the Atomic Energy Commission mentioned long distance transmission by facsimile systems. The cost of reproducing literature required by a library would more than cover the cost of a facsimile system, he asserted.

Electronic Searching. The possibility of developing electronic searching machines was affirmed by Philip R. Bagley of the MIT Digital Computer Laboratory. Such machines, he predicted, would search even more rapidly than punched card machines —by entirely different orders of magnitude.

It has not been established beyond reasonable doubt that punched cards will prove completely satisfactory for large files, i.e., over half a million documents, continued Mr. Bagley. But present computers will not do the job either. In fact, they would be slower than the IBM card scanner. Whirlwind I, the MIT digital computer, which Mr. Bagley termed the fastest in operation, can perform over 16,000 comparisons per second. But because the machine is limited to scanning for one item at a time, it would take it more than 800 hours to search the index for a million documents.

Punched card techniques, however, are limited by the speed at which cardboard can be pushed through machines. The limit on tape methods is much higher. Also, tape can carry its record in condensed form. A punched card record about 3 inches wide can be expressed as magnetized spots on a 1-inch tape.

Present computors could be rewired to improve their searching operation, but they would still scan sequentially, and they would contain many parts that would be idle during use as searching machines. Much better, said Mr. Bagley, would be to design a special-purpose scanning machine in which an index would be compared against all the searching criteria simultaneously—as in the IBM card scanner.

Such a machine would be much simpler than an all-purpose machine; it would consist of a comparator, for scanning, and a logical computer. The purpose of the computer would be for establishing that the specified relationship exists between the index entries detected by the comparator. About 800 vacuum tubes would be required in the machine, of which more than 500 would be required for the comparator unit. A magnetic drum could be used for temporary storage of the output until it could be printed either by an electric typewriter or by a device that burns characters on a paper tape.

Such an electronic searching machine, according to Mr. Bagley, could be designed to scan the index for 5 million documents per hour. At this rate it would scan the index for all the documents represented in *Chemical Abstracts*—from its founding in 1907 to date—in about 15 minutes.

PRINTED IN U. S. A.